****

**Natural Language Processing & Applications**

**21EC4082**

STUDENT ID: ACADEMIC YEAR: 2023-24

STUDENT NAME:

Table of Contents

1. **Session 01: Introductory SessionNA**
2. **Session 02: Tokenization\_of\_text #1#**
3. **Session 03: Text\_2\_Sequences #2#**
4. **Session 04: One\_Hot\_Encoding #3#**
5. **Session 05: Vectorization\_of\_texts #4#**
6. **Session 06: Databases\_how\_to\_Use #5#**
7. **Session 07: Parsing\_nltk\_toolbox #6#**
8. **Session 08: TF\_Testing\_fail #7#**
9. **Session 09: IDF\_Why #8#**
10. **Session 10: TFIDF\_Vertorization #9#**
11. **Session 11: TF\_IDF\_Failure\_meaning #10#**
12. **Session 12: Distance\_Metrics #11#**
13. **Session 13: Word\_similarities\_nltk #12#**
14. **Session 14: Document\_recognition\_tfidf\_vectors #13 (Adv/Peer)#**
15. **Session 15: Zipf's\_Law\_nlp #14 (Adv/Peer)#**
16. **Session 16: Simple\_topic\_modelling\_ex #15 (Adv/Peer)#**
17. **Session 17: PCA\_From\_SCratch #16 (Adv/Peer)#**
18. **Session 18: Singular\_Value\_Decomposition\_SVD\_Ex #17 (Adv/Peer)#**
19. **Session 19: Latent\_Semantic\_Analysis\_SVD #18 (Adv/Peer)#**
20. **Session 20: spam\_dect\_class #19 (Adv/Peer)#**
21. **Session 21: Sentiment\_Analysis\_RNN #20 (Adv/Peer)#**

<https://github.com/pvvkishore/NLP-A_LAB_2023>

**A.Y. 2023-24 LAB/SKILL CONTINUOUS EVALUATION**

| **S.No** | **Date** | **Experiment Name** | **Pre-Lab (10M)** | **In-Lab (25M)** | | | **Post-Lab**  **(10M)** | **Viva Voce (5M)** | **Total**  **(50M)** | **Faculty Signature** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Program/ Procedure (5M)** | **Data and Results (10M)** | **Analysis & Inference (10M)** |
| **1.** |  | **Introductory Session** | **-NA-** | | | | | | | |
| **2.** |  | **Tokenization\_of\_text #1** |  |  |  |  |  |  |  |  |
| **3.** |  | **Text\_2\_Sequences #2** |  |  |  |  |  |  |  |  |
| **4.** |  | **One\_Hot\_Encoding #3** |  |  |  |  |  |  |  |  |
| **5.** |  | **Vectorization\_of\_texts #4** |  |  |  |  |  |  |  |  |
| **6.** |  | **Databases\_how\_to\_Use #5** |  |  |  |  |  |  |  |  |
| **7.** |  | **Parsing\_nltk\_toolbox #6** |  |  |  |  |  |  |  |  |
| **8.** |  | **TF\_Testing\_fail #7** |  |  |  |  |  |  |  |  |
| **9.** |  | **IDF\_Why #8** |  |  |  |  |  |  |  |  |
| **10.** |  | **TFIDF\_Vertorization #9** |  |  |  |  |  |  |  |  |
| **11.** |  | **TF\_IDF\_Failure\_meaning #10** |  |  |  |  |  |  |  |  |
| **12** |  | **Distance\_Metrics #11** |  |  |  |  |  |  |  |  |
| **13.** |  | **Word\_similarities\_nltk #12** |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
| **14.** |  | **Document\_recognition\_tfidf\_vectors #13 (Adv/Peer)** |  |  |  |  |  |  |  |  |
| **15.** |  | **Zipf's\_Law\_nlp #14 (Adv/Peer)** |  |  |  |  |  |  |  |  |
| **16.** |  | **Simple\_topic\_modelling\_ex #15 (Adv/Peer)** |  |  |  |  |  |  |  |  |
| **17.** |  | **PCA\_From\_SCratch #16 (Adv/Peer)** |  |  |  |  |  |  |  |  |
| **18.** |  | **Singular\_Value\_Decomposition\_SVD\_Ex #17 (Adv/Peer)** |  |  |  |  |  |  |  |  |
| **19.** |  | **Latent\_Semantic\_Analysis\_SVD #18 (Adv/Peer)** |  |  |  |  |  |  |  |  |
| **20.** |  | **spam\_dect\_class #19 (Adv/Peer)** |  |  |  |  |  |  |  |  |
| **21** |  | **Sentiment\_Analysis\_RNN #20 (Adv/Peer)** |  |  |  |  |  |  |  |  |

**Experiment Title: Tokenization\_of\_text**

**Aim/Objective:**

**The aim is to compare and evaluate different tokenization techniques or libraries, such as NLTK, SpaCy, and TensorFlow, to determine their effectiveness in handling various types of text data.**

**Description:**

Tokenization is the 1st step in any NLP model. The experiment may aim to explore how tokenization using NLTK, spaCy, and TensorFlow can be integrated into a broader NLP pipeline or used as a preprocessing step for tasks such as sentiment analysis, machine translation, named entity recognition, or text summarization. The focus is on understanding the impact of tokenization choices on downstream model performance. The experiment may aim to analyze the performance characteristics of tokenization using NLTK and TensorFlow.

**Pre-Requisites:**

Install Python 3.6 and above using.

1. <https://pip.pypa.io/en/stable/installation/>

2. <https://packaging.python.org/en/latest/tutorials/installing-packages/>

3. <https://pypi.org/project/nltk/>

4.

[https://www.tensorflow.org/install/pip](https://www.tensorflow.org/install/pip 5)

[5](https://www.tensorflow.org/install/pip 5).

[https://spacy.io/usage](https://spacy.io/usage 6)

[6](https://spacy.io/usage 6). <https://pypi.org/project/gensim/>

**Pre-Lab:**

This Section must contain at least 5 Descriptive type questions or Self-Assessment Questions which help the student to understand the Program/Experiment that must be performed in the Laboratory Session.

1. What is tokenization in the context of NLP?
2. How can you tokenize a sentence into individual words using NLTK?
3. What is the purpose of tokenizing text in NLP?
4. Name a few tokenization techniques other than word tokenization.
5. How can you tokenize a text document into sentences using NLTK?

**In-Lab:**

1. Apply tokenization methods in the NLTK library on a 5-line text data available in NLTK.
2. Apply tokenization methods in the TF library on a 5-line text data available in NLTK.
3. Draw comparisons based on text handling capabilities.

* **Procedure/Program:**

This Section is meant for the student to Write the program/Procedure for the Experiment

***(Leave at least 2-3 Pages to record the Procedure/Program)***

* **Data and Results:**

This Section is meant for the students to collect, record the results generated during the Program/Experiment execution. Include instructions on how to present the results, such as creating tables, graphs, or visualizations.

***(Leave at least 1 Page to record the results)***

* **Analysis and Inferences:**

This Section is meant for the students to analyse their data, perform calculations Include questions or prompts to encourage critical thinking and interpretation of the data

***(Leave at least 1 Page for each Program)***

**Sample VIVA-VOCE Questions (In-Lab):**

1. **What is tokenization?**
2. **According to your exp which tokenizer API is the best?**
3. **How NLTK and TensorFlow handle tokenization for different languages.**
4. **List the Metrics used to Evaluate Tokenization Techniques.**
5. **Can you tokenize multiple text documents simultaneously using TensorFlow.**

**Post-Lab:**

1. Try tokenization in the spaCy library and compare with the NLTK and Tensorflow.
2. Try tokenization on big corpus dataset given below. <https://www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter>

* **Procedure/Program:**

This Section is meant for the student to Write the program/Procedure for Experiment

***(Leave at least 2-3 Pages for each Procedure/ Program/ Solution)***

* **Data and Results:**

This Section is meant for the students to collect, record the results generated during the Program/Experiment execution. Include instructions on how to present the results, such as creating tables, graphs, or visualizations.

***(Leave at least 1 Page for recording the results)***

* **Analysis and Inferences:**

This Section is meant for the students to analyse their data, perform calculations Include questions or prompts to encourage critical thinking and interpretation of the data.

***(Leave at least 1 Page for recording the analysis and inferences)***

|  |  |
| --- | --- |
| **Evaluator Remark (if Any):** | **Marks Secured: \_\_\_\_\_out of 50** |
| **Signature of the Evaluator with Date** |

**Evaluator MUST ask Viva-voce prior to signing and posting marks for each experiment.**

**Experiment Title: Text\_2\_Sequences**

**Aim/Objective:**

**The aim is to evaluate different techniques or libraries, such as NLTK, SpaCy, and TensorFlow, to determine their effectiveness in converting text to a sequence of numbers.**

**Description:**

The objective of converting text to a sequence of numbers is a fundamental step in natural language processing (NLP) tasks. The primary goal of this conversion is to represent textual data in a numerical format that machine learning models can process effectively. To convert text to a numerical format that enables the application of machine learning and NLP techniques.

**Pre-Requisites:**

Install Python 3.6 and above using.

1. <https://pip.pypa.io/en/stable/installation/>

2. <https://packaging.python.org/en/latest/tutorials/installing-packages/>

3. <https://pypi.org/project/nltk/>

4.

[https://www.tensorflow.org/install/pip](https://www.tensorflow.org/install/pip 5)

[5](https://www.tensorflow.org/install/pip 5).

[https://spacy.io/usage](https://spacy.io/usage 6)

[6](https://spacy.io/usage 6). <https://pypi.org/project/gensim/>

**Pre-Lab:**

This Section must contain at least 5 Descriptive type questions or Self-Assessment Questions which help the student to understand the Program/Experiment that must be performed in the Laboratory Session.

1. Why convert text to numbers?
2. How effective is the method used by you?
3. Are all sentences in the text considered to have the same length? If No, What did you do.
4. In NLTK, which function is used to assign numeric IDs to tokens?
5. What is the difference between word tokenization and sentence tokenization?

**In-Lab:**

1. Apply tokenization and convert a sequence of sentences in the NLTK library to a sequence of numbers.
2. Convert a 10-sentence dataset with multiple-length sentences into a number array of equal size for ML model training.

* **Procedure/Program:**

This Section is meant for the student to Write the program/Procedure for the Experiment

***(Leave at least 2-3 Pages to record the Procedure/Program)***

* **Data and Results:**

This Section is meant for the students to collect, record the results generated during the Program/Experiment execution. Include instructions on how to present the results, such as creating tables, graphs, or visualizations.

***(Leave at least 1 Page to record the results)***

* **Analysis and Inferences:**

This Section is meant for the students to analyse their data, perform calculations Include questions or prompts to encourage critical thinking and interpretation of the data

***(Leave at least 1 Page for each Program)***

**Sample VIVA-VOCE Questions (In-Lab):**

1. **What does NLTK's FreqDist class provide?**
2. **According to your exp which API is the best?**
3. **Do you think your sequence conversion is suitable for GPT.**
4. **List the Metrics used to Evaluate sequence conversion Techniques.**
5. **Can you convert using spaCy.**

**Post-Lab:**

1. Try normalization of converted numbers from text data.
2. Try text to sequences on big corpus dataset given below. <https://www.kaggle.com/datasets/thoughtvector/customer-support-on-twitter>

* **Procedure/Program:**

This Section is meant for the student to Write the program/Procedure for the Experiment

***(Leave at least 2-3 Pages for each Procedure/ Program/ Solution)***

* **Data and Results:**

This Section is meant for the students to collect, record the results generated during the Program/Experiment execution. Include instructions on how to present the results, such as creating tables, graphs, or visualizations.

***(Leave at least 1 Page for recording the results)***

* **Analysis and Inferences:**

This Section is meant for the students to analyse their data, perform calculations Include questions or prompts to encourage critical thinking and interpretation of the data.

***(Leave at least 1 Page for recording the analysis and inferences)***

|  |  |
| --- | --- |
| **Evaluator Remark (if Any):** | **Marks Secured: \_\_\_\_\_out of 50** |
| **Signature of the Evaluator with Date** |

**Evaluator MUST ask Viva-voce prior to signing and posting marks for each experiment.**